Junseok Park

Cambridge, MA 02141 • junseok.park@childrens.harvard.edu • junseokpark.github.io

A machine learning scientist, data scientist, or bioinformatician in computational genomics with expertise in computer science, skilled in developing AI models, designing large-scale data processing pipelines, and performing statistical analyses of next-generation sequencing and multi-omic experiments from patient cohorts—ranging from single-cell to DNA and RNA sequencing samples. Leveraging strong analytical thinking, I have research experience ranging from cancer immunology to neurodegeneration in ataxia-telangiectasia.

Skills

Tool developments: rTea, GenomeFlow, CORUS, CLIPS
Computational skills: Deep learning, Machine learning, Statistical and Network Analysis
Deeplearning frameworks: TensorFlow, Pytorch and Keras
Programming languages: Python, R, Scala, C++, Java, JavaScript, Shell Scripts, SQL and noSQL
Computing platforms: On-premises HPC (Slurm/Qsub), AWS, GCP and Terra.bio
Workflow development: WDL, Snakemake, Airflow and Kubernetes

Work Experience

Aug. 2020 – present **Lee Lab** – Boston Children's Hospital and Harvard Medical School, Boston, MA, USA
Research Fellow

- Developed a tool (rTea) to detect transposon-fusions, analyzing 13,345 RNA-seq from normal and 34 cancer types. Identified 307,793 cancer-specific fusions with 30 events per cancer.
- Employed a Generalized Linear Mixed Model to correct for six technical variables in RNA-seq data from 28 human tissue types, identifying 48% of fusions (avg 639 per sample) in testis.
- Developed GenomeFlow to process large datasets on GCP, achieving a 77% cost reduction (saving $86K) in its first application compared to the initial setup.
- Adopted Convolutional Neural Networks (CNNs) to detect somatic L1 insertions, achieving 0.885 precision and 0.845 recall from bulk WGS of a cancer cell line.
- Used SVM to validate the RF model for utilizing zebrafish as a model of Duchenne Muscular Dystrophy, showing similar outcomes (AUROC = 0.99) to the RF model.
- Customized the DRAGEN-GATK pipeline on GCP to ensure variant quality from WGS samples, supporting the identification of ATM gene mutations in six ataxia-telangiectasia cases.

Mar. 2014 – Mar 2020 **KAIST - Bio-Synergy Research Center(BSRC)** – Daejeon, Republic of Korea
Researcher - MSc/PhD student

- Developed CORUS on AWS to collect participatory trial data and verified the data reliability weight index using an independent samples t-test ($p < 0.011$) with data from 340 participants.
- Built a clinical trial protocol database system (CLIPS) from 184,634 trials, achieving a 35% higher F1-score (0.52) and 82% greater user satisfaction (6.5) than keyword search (2.3).
- Developed a Doc2Vec model with 998,543 definitions from UMLS2015 and Wikipedia, improving biomedical term similarity by 35% and coverage by 4.77% over the Vector method.
- Supported the identification of drugs with opposing effects on disease genes by combining 2,434 drug-target interactions, 166 pathways, and 364 disease genes for network analysis.

Jan. 2012 – Feb 2014 **Korea Research Institute of Chemical Technology (KRICT)** – Daejeon, Republic of Korea
Technical Manager

- Developed a chemical reagent management system to optimize reagent usage workflows.

Nov 2011 – Dec. 2011 **Seoul National University Hospital** – Seoul, Republic of Korea
Planning Manager

- Integrated medical records (EMR/EHR/PHR) to develop a new hospital management system

Apr. 2011 – Nov. 2011 **Korea BIO-IT Foundry Gwang-ju Centre** – Gwangju, Republic of Korea
Researcher (PI: Dr. Yonggwan Won)

- Designed and developed a high-speed additive reagent injection and inspection system

Jan. 2007 – Mar. 2010 **SKTelecom** – Gwangju/Seoul, Republic of Korea
Manager

- Worked independently with sub-contractors to optimize and enhance the performance of WCDMA networks, improving reliability and benefiting approximately 12 million customers.

Mar. 2001 – Apr. 2003 **Republic of Korea Army** – 31 Division, Republic of Korea
Sergeant, Honorable discharge

- Developed ocean protection radar systems and networks along the South Korean coast.

Education

2016–2020 **PhD** program in **Bioinformatics** – KAIST, Daejeon, South Korea
2014–2016 **MSc** in **Bioinformatics** – KAIST, Daejeon, South Korea
2000–2007 **BSc** in Computer Engineering – Chonnam National University, Gwangju, South Korea

Highlighted Publications

- Boram Lee*, **Junseok Park***, Adam Voshall, Yangmin Gan, Eduardo Maury, Yeeok Kang et al. *Pan-cancer analysis reveals multifaceted roles of retrotransposon-fusion RNAs.* **Nature Communication** (2025): In revision. 10.1101/2023.10.16.562422, github.com/junseokpark/rtea
- **Junseok Park**, Eduardo Maury, Changhoon Oh, Donghoon Shin, Danielle Denisko, Eunjung Alice Lee. *Genomic data processing with GenomeFlow.* Accepted to **BMC Bioinformatics** (2024). github.com/junseokpark/genomeflow
- Miaomiao Tan*, Zhinan Lin*, Zhuofu Chen, **Junseok Park**, Ziting He, Haonan Zhou et al. *Image-based DNA Sequencing Encoding for Detecting Low-Mosaicism Somatic Mobile Element Insertions.* **Nature Communication** (2024): In revision. 10.1101/2024.11.07.619809
- Jeffrey J. Widrick, Matthias R. Lambert, Felipe de Souza Leite, Youngsook Lucy Jung, **Junseok Park**, James R. Conner et al. *Kinematic phenotyping of dystrophic zebrafish larvae*, Science advances (2024): In revision. 10.1101/2024.12.05.627004
- Lai, Jenny, Didem Demirbas, Junho Kim, Ailsa M. Jeffries, Allie Tolles, **Junseok Park** et al. *ATM-deficiency-induced microglial activation promotes neurodegeneration in ataxia-telangiectasia.* **Cell reports** (2024): 43, no. 1. 10.1016/j.celrep.2023.113622
- **Junseok Park**, Kwangmin Kim, Seongkuk Park, Woochang Hwang, Sunyong Yoo, Gwansu Yi, Doheon Lee. *An interactive retrieval system for clinical trial studies with context-dependent protocol elements.* PloS one (2020): 15.9:e0238290. 10.1371/journal.pone.0238290, github.com/junseokpark/clips
- **Junseok Park**, Seongkuk Park, Kwangmin Kim, Gwangmin Kim, Jaegyun Jung, Sunyong Yoo et al. *Reliable Data Collection in Participatory Trials to Assess Digital Healthcare Applications.* **IEEE Access** (2020): 79472-79490. 10.1109/ACCESS.2020.2985122, github.com/junseokpark/corus
- **Junseok Park**, Kwangmin Kim, Woochang Hwang, and Doheon Lee. *Concept embedding to measure semantic relatedness for biomedical information ontologies.* Journal of biomedical informatics (2019): 94:103182. 10.1016/j.jbi.2019.103182

Patents

Jan 2022 Method and Apparatus for Performance Evaluating of Healthcare Applications
KR20210081545A, https://patents.google.com/patent/KR20210081545A

Jan 2020 Method and Apparatus for Data Managing for Clinical Trial
KR20190094729A, https://patents.google.com/patent/KR20190094729A

Jul 2017 A Method for Searching Co-Occurrence Based on Co-Operational Formation
KR20160149619A, https://patents.google.com/patent/KR20160149619A

Selected Awards

2021–2024 Basic Science Research Funding; 135,000,000 KRW (45,000,000 KRW annually over three years)
National Research Fund, Republic of Korea

2022 AnVIL Cloud Credits Continued Program; $10,000 USD
National Human Genome Research Institute, USA

Selected Certificates

2011–Present **Network Management** – Information & Communication Qualification Association of Korea

2011–Present **Engineer Information Processing** – Human Resources Development Service of Korea

*Co-first authors